Engineering Socially-Oriented Autonomous Agents and Multiagent Systems

Nieves Montes
Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona – Spain
nmontes@iiia.csic.es

Abstract

The emergent field of social AI is concerned with the development of autonomous agents that are able to act as part of larger community. Within this context, my research seeks to engineer meaningful social interactions among a group of agents from two different approaches. First, the societal level leverages constructs that apply to a society as a whole, like norms and social values. Second, the individual level endows agents with the ability to reason about others, by making use of Theory of Mind capabilities.

1 Overview

The main focus of my research is the study of the foundations and implementation of socially-aware autonomous agents and multiagent societies. This is included as part of the emergent field of *social AI*, which seeks to develop autonomous agents that are able to act as part of a larger community, possibly including humans. In particular, my research is concerned with: (1) the engineering of social interactions at the *societal or multiagent (MAS) level*; and (2) the development of the cognitive machinery for socially-aware agents at the *individual level*.

2 Engineering Interactions at the MAS Level

In the first part of my research, I have focused on AI techniques to engineer social interactions at the MAS level. Here, it is necessary to construct an overall model of the interaction taking place, taking into account the *norms* in place, and analyzing the most likely outcomes in terms of the *values* that the society is expected to abide by.

Following this line of research, I have proposed a general methodology for the automatic synthesis of prescriptive norms based on their degree of alignment with respect to some value [Montes and Sierra, 2021]. In this work, norms are tied to numerical optimizable parameters, which allow to use off-the-shelf metaheuristic techniques to find the set of norms (or normative systems) that most successfully promote some value of choice. Moreover, that work also provides a kit of analytic tools to examine the resulting optimal normative systems: the Shapley values of individual norms (which quantify the contribution of a single norm towards the alignment), and the compatibility of values (which quantifies to what degree the aggressive promotion of value v_i hinders the achievement of value v_j).

Despite the progress made in [Montes and Sierra, 2021], its rigid representation of norms requires to define from scratch the space of norms to search every time one wants to use the methodology in a different scenario. To tackle this limitation, in another piece of research I have defined the Action Situation Language (ASL) [Montes *et al.*, 2021], inspired by Elinor Ostrom's Institutional Analysis and Development framework [Ostrom, 2005].

The ASL allows agents to represent a wide variety of norms in a machine-readable and syntactically-friendly way (as `if-then-where` statements) and automatically assess the impact they would have in their communities if they were to be adopted, by performing *what-if* analysis. The ASL is complemented by a game engine, that takes in an action situation description and automatically builds its formal semantics as an extensive-form game. This game model, then, can be analyzed using standard game-theoretical tools. With ASL, we have been able to model several benchmark social scenarios from the policy analysis literature.

The ASL follows in the footsteps of previous languages for the systematic definition of games [Koller and Pfeffer, 1997; Schiffel and Thielscher, 2014]. However, the main feature that sets ASL apart is the fact that ASL descriptions are meant to be *extensible*. Its full power is leveraged when the effects of adding, retracting, or changing the priority of rules (which indicates the precedence of a rule statement when conflicts arise) are assessed in an automated fashion.

3 Engineering Interactions at the Individual Level

In the second phase of my research, I am focusing on the cognitive machinery that an agent must individually possess in order to reason about others. The cognitive ability to put oneself in the shoes of someone else and reason from their perspective is called *Theory of Mind* (ToM). ToM techniques are fairly prevalent in competitive domains where agents have diverging interests [Nashed and Zilberstein, 2022]. However, the incorporation of techniques for modeling others has the potential to boost the performance of agents not only in competitive settings, but in collaborative ones too.

Recently, the game of Hanabi has emerged as the perfect test bed to evaluate the performance of new techniques for modeling others in cooperative tasks. Hanabi is an award-winning card game where agents must collaborate to build stacks of cards with identical color, however they can only see the cards of others and not their own. Players can share information with one another through hints, which are quantified through a finite number of information tokens.

There are several features of Hanabi that make it an excellent benchmark to test techniques for modeling others. First, Hanabi is a purely cooperative game where agents all share a common goal and need to coordinate as a team to achieve it. Second, agents have to deal with imperfect information, as they do not have access to their own cards. Third, information itself is collectively managed by the team as a common resource. All of these features have led some researches to propose Hanabi as the next major challenge to be undertaken by the AI community [Bard *et al.*, 2020].

The approach that I am exploring for Hanabi (and for cooperative domains more generally) involves the combination of ToM techniques with abductive reasoning. The main assumption of this approach is the shared *common expertise* of all agents in the team, meaning that they approach the task endowed with the same logic theory. Importantly, agents are also assumed to have agreed *a priori* on a strategy to follow during game play. Hence, my work falls within the category of rule-based approaches, as opposed to reinforcement learning ones [Siu *et al.*, 2021].

The basic agent model consists of, first, an observer agent i getting notified that an acting agent j has selected and is about to perform some action a_j. Second, the observer i engages in ToM by simulating the perspective of the actor j. This perspective is incomplete, as i can, in general, only construct an approximation of the view that j has on the state of the system. Third, the observer i computes, through abductive reasoning, the explanations that would justify j selecting a_j. These explanations contain additional knowledge that the observer i incorporates into their own knowledge base, to make use of them for later decision-making.

Although I have been using Hanabi as the test to evaluate the performance of this agent model, it is worth noting that the cognitive machinery combining ToM and abduction seeks to be domain-independent. Note also that the cooperation model presented above is agnostic with respect to the particular strategy being implemented, as long as it is shared by all teammates.

4 Conclusions and Future Research

In summary, my research deals with the engineering of social interactions in autonomous agents, both at the collective level (through norms and values) and the individual level (by endowing agents with Theory of Mind capabilities).

In the next steps of my research, I intend to continue exploring the possibilities of the ToM+abduction agent model presented in Section 3. The first option is to extend it beyond first-order ToM and allow for epistemic strategies. Second, since the model is agnostic with respect to the particular strategy being implemented, the space of possible strategies can be automatically searched in order to find the optimal one. This direction would be in line with the optimization approach of the first contribution presented in Section 2. The third option, and the one that goes the furthest, consists of extending the ToM+abduction cognitive machinery for *ad hoc* teamwork, where agents must coordinate but do not share a pre-agreed upon strategy.

Acknowledgments

I would like to express my gratitude to my supervisors, Prof Carles Sierra and Dr Nardine Osman, for their continued support and guidance. I would also like to thank the EU TAILOR project (H2020 #952215) for their financial support.

References

[Bard *et al.*, 2020] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216, 2020.

[Koller and Pfeffer, 1997] Daphne Koller and Avi Pfeffer. Representations and solutions for game-theoretic problems. *Artif. Intell.*, 94(1–2):167–215, jul 1997.

[Montes and Sierra, 2021] Nieves Montes and Carles Sierra. Value-guided synthesis of parametric normative systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, page 907–915, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. (Best paper award finalist).

[Montes *et al.*, 2021] Nieves Montes, Nardine Osman, and Carles Sierra. *Enabling Game-Theoretical Analysis of Social Rules*, volume 339 of *Frontiers in Artificial Intelligence and Applications*, pages 90–99. IOS Press, 2021.

[Nashed and Zilberstein, 2022] Samer Nashed and Shlomo Zilberstein. A survey of opponent modeling in adversarial domains. *Journal of Artificial Intelligence Research*, 73:277–327, 2022.

[Ostrom, 2005] Elinor Ostrom. *Understanding Institutional Diversity*. Princeton University Press, 2005.

[Schiffel and Thielscher, 2014] S. Schiffel and M. Thielscher. Representing and reasoning about the rules of general games with imperfect information. *Journal of Artificial Intelligence Research*, 49:171–206, 2014.

[Siu *et al.*, 2021] Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. Evaluation of human-AI teams for learned and rule-based agents in Hanabi. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16183–16195. Curran Associates, Inc., 2021.